www.icgst.com

Text lines Segmentation of Handwritten Arabic Script using Outer Isothetic Cover

Samia Snoussi, Fethi Ghazouani, Yosra Wahabi
Faculty of Computing and Information Technology Jeddah University, Saudi Arabia
RIADI Laboratory, ENSI, Manouba university, Tunisia
College of Community Service and Continuing Education, Umm Al Qura University, Saudi Arabia
Samia.maddouri@enit.rnu.tn, gfethi@yahoo.fr, yosra_w@yahoo.fr

Abstract

In this paper, we present a new bottom up segmentation method of handwritten Arabic texts into text lines. The proposed method uses the algorithm for construction of the Outer Isothetic Cover (OIC) of a digital object. This technique allows the construction of polygons of all connected components on the binary image of the document. The specificity of this method is that it allows the extraction of shape of the script and not a line as a geometric form. This will reduce overlapping between successive lines. The second specificity of this method is its independency from the variability of script size in the same document. The set of extracted polygons contains big and small ones. The big polygons correspond to Pieces of Words and the small ones correspond generally to diacritic points or punctuation mark and they are identified by their perimeters after a training step. Based on these results and added information like position, upper and lower limit of the components given by the OIC, we can extract lines from handwritten Arabic text. The proposed method is tested and evaluated on a sub-set of 100 randomly chosen handwritten Arabic texts selected from KHATT Database. The obtained result attempts 74% and needs to be improved. A new database composed by more than 1000 pages of printed and handwritten documents for tutorial and guiding of HAJJ steps is created (The Hajj is the pilgrimage to Mecca). We are working on the application of the proposed segmentation method on the HAJJ document database in order to extract desired information.

Keywords: *text Segmentation, handwriting, Outer Isothetic Cover, KHATT Database, HAJJ Database.*

NOMENCLATURE

OIC	*Outer Isothetic Cover*
KHATT	*Arabic offline Handwritten Text database*
RLSA	*Run Length Smoothing Algorithm*
OCR	*optical character recognition*

1. Introduction

The segmentation is the most important phase in the applications of the pattern recognition. It allows preparing data for further processing steps like image processing, analysis and classification. In document processing field, the segmentation is essential for document recognition which it needs several steps of blocks classification, lines and words extraction from text blocks and normalization, words segmentation and features detection for each extracted information. This phase constitutes one of the major difficulties of all recognition system. Today, there is practically no reliable segmentation method for the handwritten Arabic document. The existing methods generally commit either ”Over” or ”Under” segmentation.

In this paper we propose a new method not to eliminate this segmentation problem, but we try to reduce it. We are interested in this work only on text document images. Essentially, for documents containing HAJJ and Umrah rules. The main application of this work is the segmentation of these documents. The identified segments will be transformed to text with OCR. Later, the obtained text will be analysed by a second intelligent system to extract HAJJ rules that will be used for automatic answers to questions about the HAJJ. This research was supported by the strategic Technologies Programs of the National Plan or Science, Technology and Innovation (MAARIFAH) in the Kingdom of Saudi Arabia. No: 13-INF134-05.

The segmentation of handwritten text is complicated by the variation of the interline distance and by the baselines undulation that often generates different orientations of the text. The characters in two adjacent lines may touch or overlap. This considerably complicates the text lines segmentation. In Arabic script, these situations frequently exist because of the presence of ascendant and/or descendant characters. On the other hand, the massive presence of diacritical symbols in Arabic script often generates false lines. In literature, most works of document segmentation are based on the decomposition of the image content into connected components. In this framework, we present some works of text lines segmentation in order to justify the choice of our proposed method. Then, we present a new lines segmentation method of handwritten Arabic text. Afterward, we present the database used for the evaluation of our method. Finally, we show and discuss the obtained results.

2. State of the art

Several segmentation methods for handwritten documents have been carried in the literature [21], [22]. Some of these methods are interested in text line segmentation [23], [26], others in words or sub-words segmentation and certain in characters segmentation [25]. The segmentation methods can be roughly divided into three approaches: top-down, bottom-up and hybrid. Top-down methods consider the document in its globality and a single orientation is normally looked. Bottom-up methods are based on the analyses of elements of low level of the image like the pixels and the connected components [24].

2.1. Related Works

The projection profile technique is a top-down method most used for text line extraction [1][2]. Based on this technique, Arivazhagan et al. propose in [1], a static approach to line segmentation in handwritten documents, where the initial image is partitioned in vertical strip. Then, the histogram projection of every vertical strip is calculated. The first candidate lines are extracted among the first strips and the overlapped connected components are assigned into text lines by bivariate Gaussian densities. Also, Zahour et al. propose in [2] a partial projection-based method combined with slant detection and partial contour tracing to segment historical Arabic documents into text lines. The X-Y cut algorithm [3] based on projection profile is considered as top-down method. The principle of this method consists of cutting a binary image horizontally and vertically in several strips. Nicolaou and Gatos propose in [4] a top-down technique to segment handwritten documents image into text lines by shredding their surface with local minima tracers. After blurring the initial image in the goal to enhance text line areas, they segment the images surface along several white paths. In [5], Shi and Govindaraju propose the Fuzzy Run Length Smoothing Algorithm (RLSA) as a bottom-up method for the segmentation of documents. The fuzzy RLSA measure is calculated for every pixel in the input image. A new gray- scale image is created based on the RLSA measure and the image is binarized. The text line patterns are extracted from the new image. This technique has been extended to an adaptative RLSA by Gatos et al. in the sense that additional smoothing constraints are set in regard to the geometrical properties of neighboring connected components [6]. The Hough transform technique is widely used for text lines extraction. Louloudis et al. have applied this methodology to extract lines from handwritten Greek documents. In [7] and [8] the authors have proposed a methodology based on Hough transform for text lines extraction. This method has been extended in [9] to a new line/word segmentation method. Another bottom-up method based on Artificial Intelligence was proposed in [10] by Nicolas et al. With this technique, the authors try to cluster the connected components of the document into homogeneous sets that correspond to the text lines of the document. A recent methodology makes use of adaptive local connectivity map technique for retrieving text lines from the handwritten historical manuscripts has been presented by Shi et al in [11]. In this method, the extraction of the text lines is done by a connected component collection and grouping after binarization of the input image with an adaptive binarization method. A new method proposed by Vasant Manohar et al in [12], this method involves grouping text lines segmented by a set of methods for segmentation of handwritten text lines in an undirected graph. The graph nodes correspond to connected components and the edge connecting pairs of connected components. Some important works are based on a hybrid method to extract text lines. Among these technique we can cite the method presented in [13] by Li et al. The authors use a Gaussian window and a level set method to detect text line boundaries. In [14] Bukhari et al. use the active contours (snakes) to segment handwritten text into lines. First, they apply a filter bank to smooth the input text image. The central line of text line components is then computed using ridges over the smoothed image. As a final step, active contours evolve over the ridges to segment text lines. Two unconstrained handwritten text segmentation methods are presented in [15] and [16]. The two methods propose also possibilities to apply their methods on any language. The first proposition is bottom-up method. It starts from gray values of pixels. However, the second method is top-down. It starts from overlapping blocks and use a skew blocs estimation.

2.2. Discussion of existent methods

Some document segmentation works estimation are based on projection [1], [2], [3], [4]. This method is very sensitive to slant, so needs some preprocessing steps such slant detection. The elimination of the slant can add distortions to the document content. Other works use some image processing tools such as binarisation, smoothing and other specific algorithms [6], [14], [15]. Third set of document segmentation methods use Hough transform such as [7], [8], [9]. The main problem of these methods are also slant. Other variety of methods based on structural description [11], [12], Artificial intelligence[9]. They try to combine image processing and pattern recognition tools [5], advanced algorithm, intelligent systems and some heuristics. The proposed method is situated in this field. It combines image processing tool and an advanced algorithm which is insensitive to slant and to size script variation.

3. Proposed Method

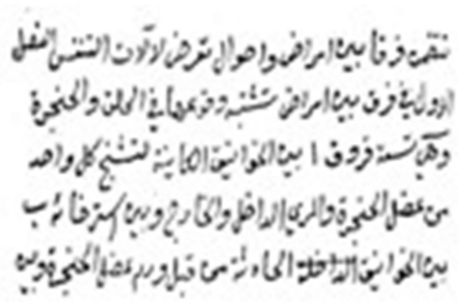
The proposed methodology for text line segmentation in handwritten Arabic document uses the algorithm of construction of the Outer Isothetic Cover of a digital object (OIC) [17]. The OIC algorithm allows constructing the polygon of each detected connected component. So, from the information relating to each polygon and the specificity of the Arabic script, we can associate each detected component to its appropriate text line. This approach has been used for the segmentation of Bengal text in words and baseline extraction by Sark in 2010 [18]. It is not used on Arabic script processing.

3.1. The Arabic Script

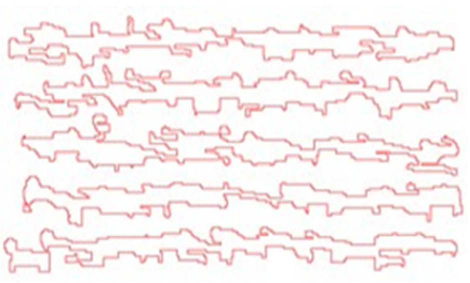
In handwritten or printed Arabic script, the Arabic word can be composed of more then one part called PAW ”Pieces of Arabic Words” or also pseudo-word. A PAW is a sequence of entirely interconnected letters. A word can be composed by one or more PAWs. Another specificity of the Arabic script is that it contains complementary signs called diacritical marks. It is a secondary component of a letter which comes to complete it or to modify the sense. It is about vowels, points or other signs (chadda, madda, hamza). The number of these points can be one, two or three and having generaly the same shape. As a result, Arabic text line may be considered as a set of ”neighbors” PAWs and diacritical points that are cited either above or below the text line. With this hypothesis and by applying the OIC algorithm on the binary image we try to associate each detected component to its appropriate text line.

3.2. The Outer Isothetic Cover: OIC

The isothetic cover of a digital object specifies a simple representation of the object and provides approximate information about its structural content and geometric characteristics. When, the cover tightly encloses the object, it is said to be an outer isothetic cover (OIC). The OIC is defined by a set of isothetic polygons, having their edges lying on the grid lines, given that the effective area corresponding to the object is minimized. The algorithm of construction of the OIC of a digital object is given by Biswas et al. in [17]. By the variation of the grid size value, the authors have presented different shapes of the OIC of the treated object. The Figure 1 shows an example of the set of outer polygons for different handwritten Arabic lines. The outer isothetic covers are obtained for grid size g=2.



(a)



(b)

Figure 1. (b) Set of polygons representing handwritten Arabic lines corresponding to results of MM of (a)

Then, to draw multiple polygons corresponding to different lines, we modify the algorithm given by [17]. This idea was used by Sarkar for words segmentation of handwritten Bangla documents in [18]. So, the algorithm is ap- plied on the result of MM on handwritten Arabic doc- ument. With a proper grid size each polygon corre- sponds to a single line in the initial document. In fact, on the result image of MM the grid points are traversed in the raw-major order until a 90 vertex. (start vertex) is found [17], Subsequent grid points are classified, marked as ”visited”, and the direction is determined from each grid point to the start vertex. Finally, the OIC is constructed when the start vertex is reached again. Figure 2 presents an example of construction of OIC for the handwritten Arabic text line after appli- cation of MM. In this figure each vertex is represented by a red point, the start vertex is surrounded (green circle) and the OIC of a line is represented with blue color.
With this process we extract text lines from a document image. In Figure 3 we show a simple example of text lines extraction from handwritten Arabic document.

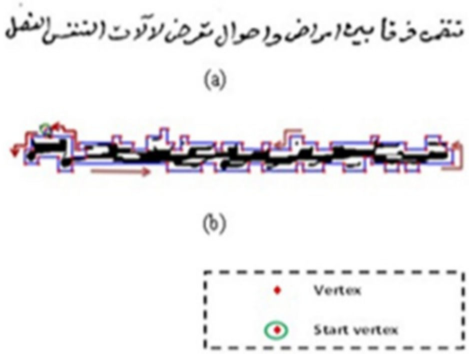


Figure 2. Construction of the OIC of handwritten Arabic lines. (a) original image, (b) result of OIC algorithm.

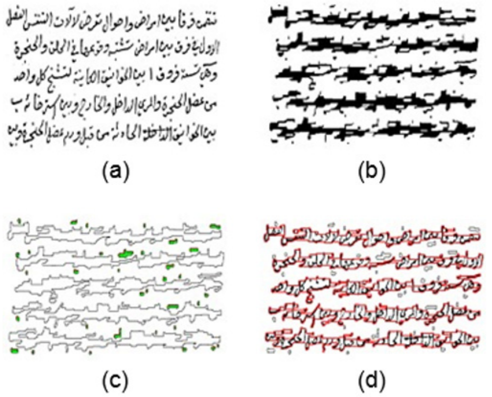
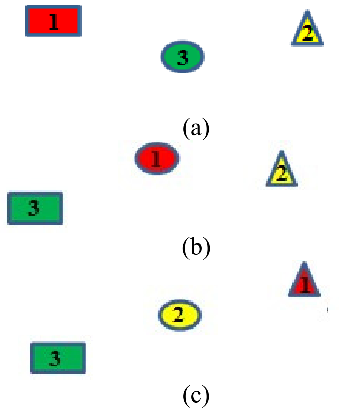


Figure 3. Text lines extraction with OIC method. (a) original image, (b) preprocessing result, (c) OIC algoritm result, (d) text lines extraction.

3.3. Construction of the Outer polygons of Components

The detection of each component in the document is done by the determination of its correspondent outer polygon. Applied on the binary image, the OIC algorithm allows constructing each polygon in the order of apparition of the component. Indeed, the algorithm traversing the grid points imposed on the binary image in the raw major order until "start vertex" is detected (the first black pixel situated in top left) [17]. Then the construction of the first polygon is determined when the start vertex is reached again. Then the algorithm determines the polygon of the second detected component by its start vertex. The procedure is iterative and it is similar to the other components in function of their apparition in the image. Figure.1 presents an example of construction of the outer polygon of three successive components. In the Figure.4 (a), the polygon of the rectangle will be constructed in the first iteration, then the triangle polygon and in the end the polygon of the circle. In the Figure.4 (b) the polygon of the circle will be constructed in the first iteration, contrary, in the Figure.4 (c) the polygon of the triangle will be the first detected.



(a)

(b)

(c)

Figure 4. Order of Construction of the Polygons

3.4. Assign Detected Component to Text Line

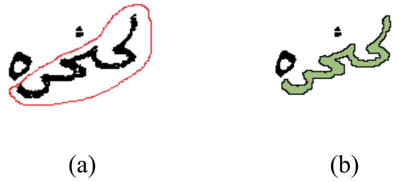
As we mentioned above, Arabic text is composed of a set of PAWs and diacritic points. In function of the perimeter of the given polygon, we can identify a PAW from a diacritic point. Generally, a diacritic point is presented by a small polygon.

On the other hand, with the detected polygon we can determine the extremities of the connected component. Indeed, the polygon is represented by its vertices of coordinates of pixel (i, j) in the image. Let $\{(x_i,y_i)\}$ (i=1..n, n: number of vertices) the set of the coordinates of vertices of the polygon. The maximum (respectively the minimum) of the $\{x_i\}$ represents the upper limit (respectively the lower limit) of a connected component. Thus, perimeter, upper and lower limit of each detected component are important information that will help us to associate the component to its appropriate text line.

3.3.1. Association of PAW to its Appropriate Text Line:

By comparing the value of perimeter of the polygon of the detected component to a value *parm* allows us to identify a PAW compared to a diacritic point. Two existing case for associate a PAW to text line. The process is as following.

- Case 1: PAW can be detected in the first iteration of the application of the OIC algorithm. In this particular case, the PAW is affected to the first text line in the treated document. The PAW will be labeled by a color C_k. Figure.5 shows an example of this case.



(a) (b)

Figure. 5. Case when The First Detected Component is a PAW

- Case 2: PAW is detected after a labeled PAW or a labeled diacritic point. In the Figure.6, the current PAW is detected after a labeled diacritic point (Figure.6 (a)), so, it will be labeled with the same color since it verifies the property (1c) (Figure.6(b)). In this case, the current PAW will be decided belonging to the line of the previously detected component or not, in function of its limits and those of its predecessor. We have proposed three properties that can verify this situation. Indeed, let $minPAW_c$ (respectively $maxPAW_c$) the upper limit (respectevely the lower limit) of the current PAW and let $minPAW_p$ (respectively $maxPAW_p$) the upper limit (respectevely the lower limit) of its predecessor detected PAW. Whether the property (1a) or (1b) or (1c) is verified, the current PAW is assigned belonging to the same text line of the predecessor PAW and it will be labeled with the color C_k of its predecessor. Subsequently the same process is done for the other successive detected PAW until the properties (1a),(1b) and (1c) are not verified. In this step, we attribute another color C_{k+1} to the PAW that does not verify those properties and we start constructing the new text line with the same principle.

Figure.7 illustrates an example for the case where the current PAW is preceded by another PAW. As shown in this figure, the first detected paw is the second in the handwritten word (Figure.7(a)) and it will be the first labeled (Figure.7(b)). Then, the third paw is detected in the second iteration of the algorithm (Figure.7(b)) and it will be labeled with the same color of its predecessor (Figure.7(c)) since it verifies the two properties (1a) and (1c).

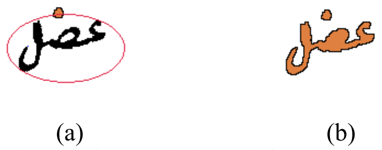

(a) (b)

Figure 6. The PAW is detected after Diacritic Point

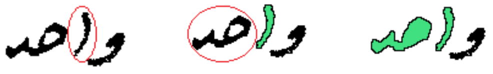

(a) (b) (c)

Figure. 7. The Predecessor of the Current PAW is another PAW

In the Figure.8, we illustrate an example of text line extraction. Figure.5(a) presents the initial image. Figure.8(b) shows the detection of the current PAW (circled in red) after the extraction of the first text line. For this current PAW neither property (1a), nor (1b), nor (1c) is verified. In this case, we attribute a new color for this PAW (Figure.8(c))and we start to construct a new text line. The final result is shown in the Figure.8(d).

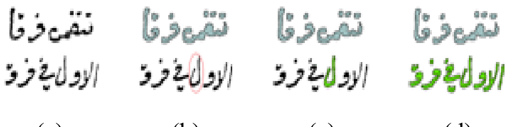

(a) (b) (c) (d)

Figure. 8. The PAW is detected after PAW

$$minPAW_p \leq minPAW_c \leq maxPAW_p \quad (1a)$$

$$|maxPAW_c - maxPAW_p| \leq S$$

$$minPAW_p \leq maxPAW_c \leq maxPAW_p \quad (1b)$$

$$|minPAW_c - minPAW_P| \leq S \quad (1c)$$

Where the threshold value S is selected and fixed so that the current PAW and its previous are nearest in the sense of their limits.

3.3.2. Association of Diacritic Point to its Appropriate Text Line: Diacritic point is a connected component that is cited above or below the PAW. On the other hand, small polygons extracted with the OIC algorithm generally correspond to the diacritic point or punctuation mark. We identify these components by comparing their perimeter to a chosen value *parm*. In our case, the value *parm* is fixed to 45 after a learning phase. Based on these two hypotheses, we try to associate diacritic point to its appropriate text line.

Diacritic point cited above PAW: two different cases can exist:

Case 1: On the first iteration, the connected component detected by the OIC algorithm is a diacritic point (Figure.9 (a)). So, the point will be affected to the first text line and will be labeled with the first color $C_k(k=1)$ (Figure.9 (b)).

(a) (b)

Figure.9. The First detected component is a Diacritic Point

Case 2: The diacritic point is detected after one or more PAWs. In this case, the point will be associated to the same text line of its previous components if it verifies the property (2), where *minPD* represents the upper limit of the diacritic point and $\{minPAW_i\}$ represents the set of the upper limits of the *i* previous PAWs. Figure.10 shows an example of this case.

$$minPD \geq \min\{minPAW_i\} \quad (2)$$

(a) (b)

Figure. 10. The upper limit of the diacritic is equal to or situated after the upper limit of the precedent paw.

•Diacritic point cited below PAW: Two properties to verify for deciding that diacritic point either belonging to same text line of its previous components or else it belonged to the next text line.

Case 1:

$$minPD \leq \max\{maxPAW_i\} \quad (3)$$

Where minPD represents the upper limit of the diacritic point and $\{maxPAW_i\}$ represents the set of the lower limits of the i previous PAWs.

Case 2: if the precedent property (3) does not satisfy, we verify, if the upper limit of the diacritic point is close to the maximum of the set of lower limits of its previous components (i.e. it verifies the property (4)).

$$|minPD - \max\{maxPAW_i\}| \leq S \quad (4)$$

Where the threshold value S is selected and fixed after a number of learning test.

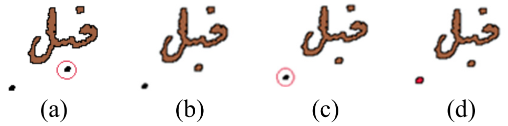

(a) (b) (c) (d)

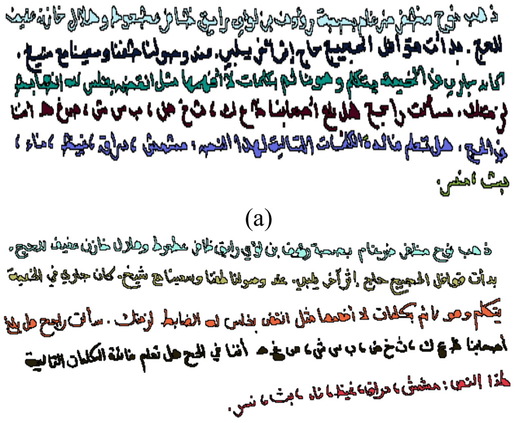
Figure. 11. Two Different Cases of Labeling of Diacritic Point

For the diacritic point cited below PAW (Figure.11(a,c)), it will be decided belonging to the same text line of the PAW if it verifies respectively (3) or (4), in this case it will be labeled with the same color C_k of the previous PAW

(Figure.11(b)). If neither (3) nor (4) is satisfied, the point will be labeled with the color C_{k+1} of the next text line (Figure.11(d)).

4. Experimental results

The proposed text line extraction method is tested on a set of 100 randomly chosen handwritten Arabic text images selected from KHATT Database [20].The khatt Database is a collection of 1000 handwritten forms written by 1000 distinct writers from different countries. The database contains also 2000 randomly selected text paragraphs from 46 sources, 2000 minimal text paragraph covering all the shapes of Arabic characters, and optionally written paragraphs on open subjects. The 2000 random text paragraphs consist of 9327 lines. The database forms were randomly divided into 70%, 15%, and 15% sets for training, testing, and verification, respectively. The sub-set of the 100 handwritten Arabic text images are chosen from the set ”Test” of KHATT Database. This sub-set text images are written by different writers. Some of them have variable skew angles among text lines. In addition, there are text images having text lines with different skew directions as well as text images having text lines with converse skew angles along the same text line. The total number of text lines is 590. Manual evaluation of the results of our method on these text images shows a correct extraction rate of 74%. Figure.12 illustrate an example of correct text line extraction for a text written by different writers.



(a)

(b)

Figure. 12. Correct Text Line Extraction. (a) Text with small interline and (b) Text with skewed and curved text lines

The weak correct extraction rate is explained by different factors. Indeed, it might be due to the choice of the two threshold S and *parm* for associating connected components to their line. the scecond factor is explained by the fact that a component of a line $(i+1)$ may be detected before some components of the line (i), this case is almost presents in case of two adjacent curved or skewed lines. Another factor is due to the touching and overlapping of certain lines. Finally, it can be explained by the small number of the used text images for the evaluation. Figure.13 shows an example of incorrect text line extraction. This work is actually evaluated on the HAJJ database [27] created to improve our actual research on the innovated integrated intelligent and mobile system for tutorial and guiding of HAJJ steps based on geolocalization.

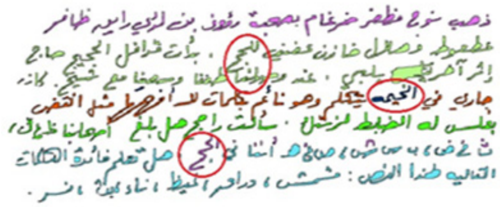


Figure. 13. Incorrect Text Line Extraction

5. Comparison with existent methods

The comparison with existent methods is done essentially on the size of data evaluation. The obtained results can be compared to presented method only if the evaluation is done on the same datasets and by the use of the same evaluation method. In table II, we present some obtained results.

Most of these methods do not indicate the origin of their document. Even if they indicate the name of the used database, they do not indicate how they choice the training set and the testing set. They only give an idea about the size of their dataset. The proposed method gives the worst segmentation rate . Firstly, we should do comparison on the same dataset. Secondly, improvement of the segmentation method and or the evaluation method is also needed.

TABLE II. SEGMENTATION METHODS COMPARISON

Approaches	Database Language	Pages	Line number	Segmen-tation rate
Arivazhagan [1]	English and **Arabic**	720	11581	97.31
Zahour [2]	Printed or handwritten historical **Arabic**	100		96
Nicolaou [4]	Latin	80	1771	98.8
Louloudis [10]	Latin	152	3382	95.8
Shi and al [14]	DARPA MADCAT database *	45	1022	99.5
Alaei et. Al. [18]	Persian text	52	823	92.35
OIC approach	**Unconstrained Arabic** handwriting KHATT database	100	590	74

* *blank paper, paper with pre-printed ruled lines and letterheads with company logos*

6. Conclusion

We propose a new bottom-up method based on the algorithm of construction of the outer isothetic cover (OIC) for handwritten Arabic text line segmentation. By comparing respectively the upper and lower limits (respectively the perimeter) of two consecutive detected components to a threshold, the method allows to associate PAW or diacritic point to its appropriate line. This method can extracted skewed, slightly overlapping and curved text lines.

In many cases, the algorithm of the construction of the OIC can detect the components of the line i + 1 before these of line i, in this case, the components of i will be associated to the line i + 1. In the scope of feature work, we consider developing a processing step that allows solving and overcoming this problem. We will focus also to test our method on a larger number of image documents such as the HAJJ database and we will aim to evaluate the performance of the proposed method by counting the number of matches between the detected text lines by our method and the text lines in the ground truth by using the MatchScore Table.

The main perspective of this work is to be able to extract correctly information from Arabic scanned documents. The extracted information will be transformed to text with OCR. Later, the obtained text will be analysed by an intelligent system to extract HAJJ rules that will be integrated to a mobile system for tutorial and guiding of HAJJ steps based on geolocalization (GPS).

7. Acknowledgements

This research was supported by the strategic Technologies Programs of the National Plan or Science, Technology and Innovation (MAARIFAH) in the Kingdom of Saudi Arabia. No: 13-INF134-05.

8. References

[1] M. Arivazhagan, H. Srinivasan, and S. N. Srihari, “A statistical approach to handwritten line segmentation,” in Document Recognition and Retrieval 14, Proc. of SPIE, San Jose CA, USA, pp. 65 000T.1– 65 000T.11, 2007.

[2] A. Zahour, L. Likforman-Sulem, W. Boussalaa, and B. Taconet, “Text line segmentation of historical arabic documents,” in Proc. of the International Conference on Document Analysis and Recognition (ICDAR), Curitiba, Paran, Brazil, pp. 138–142, 2007.

[3] F. Shafait, D. Keysers, and T. M.Breul, “Performance comparison of six algorithms for page segmentation,” IAPR Workshop on Document Analysis Systems, vol. 3872, no. 1, pp. 368–379, 2006.

[4] A. Nicolaou and B. Gatos, “Handwritten text line segmentation by shredding text into its lines,” in Proc. of the the International Con- ference on Document Analysis and Recognition, Barcelona, Spain, pp. 626–630, 2009.

[5] Z. Shi and V. Govindaraju, “Line separation for complex document images using fuzzy runlength,” in Proc. of the International Workshop on Document Image Analysis for Libraries (DIAL), Washington, DC, USA, p. 306, 2004.

[6] B.Gatos, A. Antonacopoulos, and N. Stamatopoulos, “Line separation for complex document images using fuzzy runlength” in Proc. of the International Conferenceon Document Analysis and Recognition (ICDAR), Curitiba, Brazil, pp. 1284–1288, 2007.

[7] Fadoua Bouafif, Samia Snoussi Maddouri, Noureddine Ellouze, On Segmentation Methods for multilingual and mixed Scripts, ICGST-GVIP journal emphasizes on graphics, vision and image processing, 2009.

[8] G. Louloudis, B.Gatos, and I.Pratikakis, “Text line detection in hand- written documents,” PATTERN RECOGNITION, vol. 41, pp. 3758–3772, 2008.

[9] G. Louloudis et al., “Text line and word segmentation of handwritten documents,” PATTERN RECOGNITION, vol. 42, pp. 3169–3183, 2009.

[10] S. Nicolas, T. Paquet, and L. Heutte, “Text line segmentation in handwritten document using aproduction system,” in Proc. of the Inter- national Workshop on Frontiers in Handwriting Recognition (IWFHR), Tokyo, Japan, pp. 245–250, 2004.

[11] Z. Shi, S. Setlur, and V. Govindaraju, “A steerable directional local profile technique for extraction of handwritten arabic text lines,” in Proc. of the International Conference on Document Analysis and Recognition (ICDAR), Barcelona, Spain, pp. 176–180, 2009.

[12] V. Manohar, S. N. Vitaladevuni, C. H. Cao, R. Prasad, and P. Natarajan, “Graph clustering-based ensemble method for handwritten text line segmentation,” in Proc. of the International Conference on Document Analysis and Recognition (ICDAR), Beijin9, China, pp. 574–578, 2011.

[13] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger, “Script-independent text line segmentation in freestyle handwritten documents,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, pp. 1313–1329, 2008.

[14] S.S. Bukhari, S. F., and T. M. Breuel, “Script-independent handwritten textlines segmentation using active contours,” in Proc. of the International Conference on Document Analysis and Recognition (ICDAR), Barcelona, Spain, pp. 446–450, 2009.

[15] A. Alaei, P. Nagabhushan, Umapada Pal: Piece-wise painting technique for line segmentation of unconstrained handwritten text: a specific study with Persian text documents. Pattern Anal. Appl. 14(4): 381-394, 2011.

[16] M. Ziaratban, K. Faez: Adaptive Script-Independent Text Line Extraction. IEICE Transactions 94-D(4): 866-877, 2011.

[17] A. Biswas, P. Bhowmick, and B. Bhattacharya, “Construction of iso- thetic covers of a digital object: A combinatorial approach,” Journal of Visual Communication and Image Representation, vol. 21, pp. 295–310, 2010.

[18] A. Sakar, A. Biswas, PBhowmick, P., and Bhattacharya, B. (2010). Word segmentation and baseline detection in handwritten documents using isothetic covers. In ICFHR10, pages 445–450, 2010.

[19] S. A. Mahmoud, I. Ahmad, M. Alshayeb, W. G. Al-Khatib, M. T. Parvez, G. A. Fink, V. Margner, and H. E. Abed, “Khatt: Arabic offline handwritten text database,” in Proc. of the International Conference on Frontiers in Handwriting Recognition (ICFHR), Bari, Italy, Sep, pp 447–452, 2012.

[20] I. Phillips and A. Chhabra, “Empirical performance evaluation of graphics recognition systems,” EEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 9, pp. 849–870, 1999.

[21] Y. Alginahi, A survey on Arabic character segmentation, International Journal on Document Analysis and Recognition (IJDAR), Volume 16, Pages: 105-126, 2013.

[22] A.Elnagar , R.Bentrcia, A Multi-Agent Approach to Arabic Handwritten Text Segmentation, International Journal of Advanced Research in Computer Science and Software Engineering, , Volume 4, Pages: 207-215, 2012.

[23] A. T. Jamal, N.Nobile, and C.Y.Suen, Shape Recognition for Arabic Handwritten Text Segmentation, Artificial Neural Networks in Pattern Recognition, Canada, Volume 8774, Pages: 228-239 2014.

[24] N. Aouadi, S. Amiri and A. Kacem Echi, Segmentation of Connected Components in Arabic Handwritten Documents, International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA), Pages: 738-746, 2013.

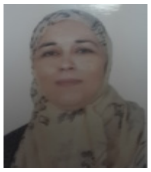
[25] A.M. Zeki, M S. Zakaria and C-Y Liong, Segmentation of Arabic Characters: A Comprehensive Survey, Technology Diffusion and Adoption: Global Complexity, Global

Innovation, chapter 16, 2013.

[26] A. T. Jamal, N.Nobile, and C.Y.Suen, Shape Recognition for Arabic Handwritten Text Segmentation, Artificial Neural Networks in Pattern Recognition, Canada, Volume 8774, Pages: 228-239 2014.

[27] Y.Wahabi, S. Snoussi Maddouri, HAJJ and Umrah digital library for the preservation and documentation of references and Arabic manuscripts, Arabic language and modern softwares conference, AlJouf University, Saudi Arabia, 2014

Biographies



Samia Snoussi Maddouri, received her diploma (Dipl.-Ing.) from the faculty of science of Tunis in 1995, Master in system and signal processing and doctorate (Dr.-Ing.) degrees in electrical engineering from the National engineering school of Tunis (ENIT), in 1997 and 2003 respectively. Since 1995, she has been working as university teacher in different Tunisien institutes. Currently she is a member of the research staff in the Image and signal processing laboratory at ENIT and teaching staff at faculty of computer science and information technologies at Jeddah university in Saudi Arabia. She worked as chef of computer science department in Taiba university. Her main areas of research are image processing and pattern recognition. Currently, she is working on Document segmentation and analysis methods and Arabic script recognition. She developed recognizer for handwritten words of Tunisian town names of IFN/ENIT database. Structural global primitives and local Fourier descriptors with a Transparent Neural Network based recognizer are the key features of her solution. Since 1998 she is also working on Document segmentation methods, at the beginning on bank checks and then on different text documents printed and handwritten. The distinction between Arabic and Latin handwritten and printed script is also one on her area of research. These Works are done in close cooperation with Braunschweig Technical University, Germany. She is one of the first developers of the IfN/ENIT-Database in 2002 (used by more than 100 research groups from more than 30 countries). She published more than 40 papers including journal papers. She was a member of DAAD cooperation project "on the Way to the Information Society: Branshweig-Tunis" during 3 year and participates to the organization of workshops in this field and represents the members of her laboratory and the Tunisian young academics women in the DAAD meeting of Project Representatives in the Programme German-Arab/Iranien Higher Education Dialogue: Dialogue through Cooperation at the Technical University, Germany in 2008. Actually I am a member of a the project intitled "Innovated Integrated Intelligent and Mobile System for tutorial and guiding of HAJJ steps based on geo-localization (GPS)".

Fethi Ghazouani, received his diploma (Master's degree in Computer Science) from the Faculty of Sciences of Tunis (FST), Tunisia in 2005 and he has received his Master degree in Computer Sciences (Data, Knowledge and Distributed Systems) from the University of Jendouba (FSJEG), Tunisia, in 2013. He has part of university study of the skills in programming, in data mining, in image processing, in artificial learning, oriented databases, etc. In his master study, he has working on Document segmentation and analysis methods and Arabic script recognition. He has developed a system for segmentation of Arabic documents that allows segmenting the document into text lines, words and characters. This work is done in cooperation with the IfN institute in Braunschweig University, Germany in 2011. Currently, he is working toward the Ph.D. degree with the Ecole Nationale des Sciences de l'Informatique (ENSI), University of Manouba, Tunisia. He is also a Permanent Researcher at RIADI Laboratory, University of Manouba, since 2014. His work is mainly related with machine learning and knowledge modeling applied to the remote sensing images.

Yosra Wahabi has the License in Applied Computer Sciences to the Management from Economic Sciences and Management faculty of Nabeul in Tunisia after presenting a research about a Database Management System of arabic document images collaborated with the Image and signal processing Laboratory of the National Engineering School in Tunis (ENIT). She has the research Master degree in signal processing with the same laboratory in 2007. Her main areas of research are image processing, object tracking and pattern recognition. Actually she is working on document segmentation, analysis methods and Arabic script recognition. In 2008 she has the certificate of skill in computer science C2i, Master Teacher from the Educational ministry in Intel program. She has research training in March 2009 with the INRIA group, Sophia Antipolis in France. In the same year she provides training workshops in computer science for teachers of other disciplines working in the educational ministry of Tunisia. She worked as computer science teacher in the same ministry for six years correlated with many trainings in the pedagogy field. Since 2012 she is a teacher in Umm Al Qura University, Saudi Arabia. She published three papers: Arabic language and modern softwares conference, Aljouf, Saudi Arabia in 2014, The International Arab Journal of Information Technology (IAJIT) in 2010 and the International Conference on Image Processing, Computer Vision, & Pattern Recognition (IPCV), USA in 2008. Since 2015 she is a member of the "Innovated Integrated Intelligent and Mobile System for tutorial and guiding of HAJJ steps based on geo-localization (GPS)" research project.